

# Judicial Performance Evaluations as Biased and Invalid Measures:

## Why the ABA Guidelines Are Not Good Enough<sup>1</sup>

3/30/2012

Rebecca Gill

University of Nevada Las Vegas

[rebecca.gill@unlv.edu](mailto:rebecca.gill@unlv.edu)

<http://ssrn.com/author=792278>

**ABSTRACT:** Judicial performance evaluations (JPEs) are an important part of the judicial selection process in the states, particularly those using a version of the merit plan. All states that use JPEs follow the ABA's Guidelines (1985), which claim to minimize the potential for unconscious bias through the use of behavior-based evaluation. But these measures have yet to be subjected to rigorous analysis. This analysis of the "Judging the Judges" survey of Nevada attorneys provides such an analysis. After controlling for objective measures of judicial performance, gender and race still contribute significantly to the scores on all of the behavior-based measures implemented in the Nevada poll. I find evidence of significant unconscious bias, as social cognition theory would predict. The analysis also cast serious doubt on the overall validity of these measures of judicial quality. This result raises serious questions about the validity and fairness of JPEs around the country.

---

<sup>1</sup> Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, April 12, 2012. A previous version of this paper was presented at the Annual Meeting of the Western Political Science Association, March 24, 2012, Portland, OR.

*In my opinion this so-called “gender bias” is purely fiction—an invention brought about by a small faction of women lawyers who are emotionally and intellectually immature. Some women want equality, but don’t want the burden of the negative effects of equality. Negative treatment at times is part and parcel of practicing law—it has absolutely nothing to do with this mythical “gender bias” which to me simply does not exist.*

(38-year old male respondent quoted in Coontz 1995, p. 16)

Judicial Performance Evaluations (JPEs) are a critical part of selecting judges, especially in states using merit-based selection systems. These JPEs, which have been increasing in popularity in recent years, fulfill two important functions. First, JPEs provide voters with information that they rely upon when casting their votes in judicial elections (Esterling 1998; Hall 1985). Second, the evaluation of judicial performance is an important tool for protecting the quality and accountability of judges (White 2009). Ideally, JPE programs should be “designed and administered in a way that does not inadvertently harm the principles they are intended to promote” (Esterling 1998, 207), especially in light of the efforts of various “good government” interest groups to implement merit-based selection systems in states around the country (Cann 2006).

But it is not clear that JPEs are designed and administered in this way. Most states follow the American Bar Association’s “Guidelines for the Evaluation of Judicial Performance” (1985). A recent update to the ABA’s Guidelines (American Bar Association 2005) acknowledges the potential for gender bias in judicial performance evaluations. But the paper dismisses the problem out of hand on the basis of the fact that the Guidelines recommend behavior-based evaluation questions:

An additional benefit of behavior-based evaluation instruments is that questionnaire items reduce subjectivity in assessments of judicial performance, thus limiting the potential for gender and other biases to influence responses. In a behavior-based evaluation instrument, items relate to judges’ actual behaviors rather than characterizations of judges’ actions as proper or improper. (American Bar Association 2005, 14).

But these guidelines and their behavior-based evaluation measures have not been validated through rigorous analysis. Research is only now beginning to address the critical question of how to minimize gender and race bias in the judicial evaluation process (Durham 2000). But, of the twenty-two states that have state-sponsored JPE systems in place, just *one* has undertaken the work of reviewing JPE surveys for empirical evidence of systematic gender and race bias,

and that study is almost 20 years old. A 1993 study of the results of the Colorado Judicial Performance Evaluation Commission's lawyer survey showed that male and female lawyers alike rated female judges consistently lower than male judges (Sterling 1993). Colorado has since adjusted its evaluation methods, but no rigorous follow up studies have been conducted to confirm that the disparities have been resolved.

### **Unconscious Bias and Judicial Performance Evaluation**

Traditionally, gender and race discrimination have been understood to be products of conscious motive or intent (Krieger 1995). But it is likely that this kind of gender- and race-related hostility makes up only a fraction of the bias we might see in professional performance evaluations. For example, male legal professionals tend to perceive much less gender bias in the workplace than do their female colleagues (Coontz 1995). This is exemplified in the introductory quote. Research also shows that, even in the context of increasing diversity initiatives on the part of law schools, race-based stereotypes of law students have a disproportionately negative effect on minority students (Clydesdale 2004). Indeed, achievement levels for minority lawyers still lags, even in the face of economic incentives for law firms to increase racial diversity (Gordon 2003).

Social science research, especially in the field of cognitive psychology, has identified a more innocent but pernicious cause of gender and race discrimination: unconscious bias. The process of simplifying and categorizing our environment, which exists is a necessary condition for most higher-level cognitive function, processes people just as it does letters, shapes, and colors (Lee 2005). Even absent a conscious bias against women or minorities, everyone is exposed to the societal stereotypes associated with different categories of people. It is through the lens of these stereotypes that we perceive, process, store, recall, and synthesize information about people. Our actions may be based in part upon the accumulated stereotypes about a particular outgroup, resulting in inaccuracy and unfairness based on race or gender.

The social science evidence for unconscious race and gender bias in employment decisions is strong and convincing. In fact, this theory of decision making played a pivotal role in the Supreme Court's decision in *Price Waterhouse v. Hopkins* [(1989) 490 U.S. 228], which held that gender stereotypes had been used to deny a female accountant's bid for partner (Fiske et al. 1991).

Social cognition theory holds that humans are naturally programmed to apply cognitive schemas to aspects of our interpersonal relationships. Just as we use situational stereotypes as shortcuts to understanding our physical world, we also develop them to organize our interpersonal interactions. This works nicely when we are aware of what we are doing and when we can control the content and activation of these schemas. But implicit social cognition theory holds that this is usually not the case; instead, we are gathering information and

categorizing people at a subconscious or unconscious level. Implicit cognition is "the process through which we become sensitive to certain regularities in the environment (1) in the absence of intention to learn about those regularities, (2) in the absence of awareness that one is learning, and (3) in such a way that the resulting knowledge is difficult to express" (Cleeremans 2003, 491). Implicit social cognition is the application of this cognitive process to information about groups of people.

This is what gives rise to unconscious bias. And this kind of bias happens much more furtively than bias based on explicit racism or sexism. Unconscious bias theory is a logical extension of implicit social cognition. People who self-report low levels of racial or gender bias can still exhibit implicit bias driven by underlying stereotype schemas (Lee 2005). This does not mean that self-reported measures of sexism and racism are disingenuous; instead, people are "unable to know the contents of their mind" (Kang and Banaji 2006, 1071), and the stereotypes creep in to frame our evaluations and behaviors of others without our conscious consent.

A few of aspects of unconscious bias theory are particularly relevant to JPEs. First, higher rates of bias tend to occur in hiring-related decisions where the characteristics that are stereotypical for the job are at odds with the gender or race stereotype (Heilman 1983). This often results in a paradox or "double bind" for women in the legal profession because they are penalized in their performance evaluations both for being too masculine and for not fitting the masculine stereotype for the job (Bowman 1998).

A second important characteristic of unconscious bias is the fact that subjective evaluation criteria exacerbate discriminatory employment decisions (Fiske et al. 1991). In JPEs, "[t]he force of traditional stereotypes is compounded by the subjectivity of performance evaluations" (Rhode 2001, 15). Previous research finds that the yes-or-no question, "Should Judge X be retained?" in Nevada's Judging the Judges survey has this effect (Gill et al. 2011). The work of judges and other legal professionals is often based at least partially on subjective assessments, "relying on the judgments of supervisors and colleagues regarding the less measurable activities" (Choi et al. 2009, 1319).

Other characteristics of the evaluation environment may also exacerbate unconscious gender and race bias. Evidence suggests that the anonymity of evaluations increases the effects of implicit bias (Hekman et al. 2010). Evaluations that are done quickly are also more subject to this kind of bias (Carnes et al. 2005). Evaluations of performance after the fact can also encourage bias, as the evaluator is required to access stored information. Information that is inconsistent with existing unconscious stereotypes is more difficult for the brain to store, but supporting evidence may be magnified in the memory—and even embellished or fabricated unknowingly (Bartlett 1932).

All of these conditions hold in attorney surveys of judicial performance. Judging is a male-stereotyped position. The types of questions asked are generally subjective. These are anonymous surveys. They are often done quickly, as attorneys are asked to rate several judges at a time on their performance over the past two years. In all, attorney surveys of judicial performance may be even more likely than other performance evaluations to suffer from unconscious gender and race bias.

These insights have important implications for assessing the evaluation process. There are increasing calls for reliance on JPEs as a way of ensuring quality standards in the judiciary (White 2009). In this context, it is imperative that JPEs not reproduce—even inadvertently—a system that disfavors groups like women and minorities, who have been historically underrepresented in the judiciary. Unfair and biased evaluations do not only harm the individuals subject to them, but they can have far-reaching and deleterious effects on the courts as an institution.

Survey research is a very complicated task, but many JPE programs have not been put together by experts in assessment methodologies. Typically, committees are made up of attorneys who lack such expertise. These committees often engage a single consultant who may or may not have all of the necessary areas of expertise (Wood Gill and Lazos 2009). To date, there has been no comprehensive assessment of the potential for unconscious race and gender bias. In the absence of this research, performance evaluation committees are forced to proceed blindly, hoping that the evaluations they are conducting do not systematically disadvantage women and minorities.

The limited evidence that we have so far indicates that this is a risky gamble. Most of the previous research underlying this cognitive bias theory has relied on self-reported feelings of bias; the research presented here provides a more systematic evaluation of gender and race based disparities in actual performance evaluations.

To date, the small amount of research that has been done on bias in judicial performance evaluations has focused on what is arguably the most subjective question on the survey: “Should Judge X be retained?” (Burger 2007; Gill et al. 2011). The stereotyping that leads to unconscious bias is exacerbated in situations where the evaluation criteria to be used are ambiguous. Certainly this is the case in the retention question. But JPEs around the country use more than just one yes/no question; all of them include a series of more specific questions intended to capture a particular dimension of judicial quality. The JPE programs currently in

existence rely heavily on the categories and questions found in the ABA Guidelines (American Bar Association 1985).<sup>2</sup>

It is likely that certain questions in the evaluation process trigger different gendered and raced understandings of what it means to perform that trait well. There is evidence that judicial temperament and legal knowledge survey questions introduce systematic gender (Durham 2000) and possibly race bias. In the 2008 “Judging the Judges” survey, male judges scored far higher than female judges on the question measuring courtesy; The highest scoring female judge scored 63%, which is lower than the two *lowest* scoring male judges (Geary 2008). As of this time, there is no published research assessing the validity of these measures. This analysis is intended to fill this gap in the research.

### **Data and Methodology**

The judicial performance evaluation data in this analysis come from the *Las Vegas Review-Journal’s* “Judging the Judges” survey of attorneys. This survey of practicing attorneys is conducted every even year, in conjunction with the judicial election cycle. I have used the publicly available aggregate data for each judge in the survey from 1998 through 2010.

The “Judging the Judges” survey asks respondent attorneys to rate the performance of the all of the judges on the Clark County ballot before whom they have had a case. The resulting data are aggregated by judge for each year, yielding 343 average scores across 93 judges. This is an unbalanced panel dataset, as not all judges served across the entire twelve year survey period. Attorneys self-select the judges they will evaluate, so the aggregated scores for each observation are based on a different number of attorney evaluations. The average is 200 attorneys, but this number ranges from 23 responses to 467.

This survey asks attorneys questions intended to measure the judicial performance qualities identified by the American Bar Association (1985). Among these qualities are legal knowledge, communication skills, integrity, administrative performance, and judicial temperament (see Table 1. Each of these qualities is measured on a three point scale, with zero being “less than adequate,” 1 being “adequate, and 2 being “more than adequate.” The aggregated data present the average of all attorney responses for each question. For those issue categories which were measured using more than one question, I have used the mean score from these questions. These mean scores are the dependent variables in the analyses, and they are outlined in the section below. Table 1 presents a summary of the relevant questions and the corresponding ABA Guidelines category. Descriptive statistics for the dependent and independent variables are summarized in Table 2.

---

<sup>2</sup> See Gill et al. (2011, at 735-36) for a table of the ABA Guidelines and the list of states that use questions measuring each of the categories and subcategories.

### *Survey Scores on Categories of Judicial Performance*

The legal ability measure is based on the ABA Guidelines category of the same name. The relevant questions ask about whether the judge's rulings are appropriate, whether the judge has reviewed the case documents, and whether the judge applies the law properly. Attorneys surveyed give the judges an average score of 1.24.

The integrity measure is based on the ABA Guidelines category called "Integrity and Impartiality." The Judging the Judges survey asks questions about a judge being free from the appearance of impropriety generally, as well as questions attempting to determine a judge's lack of bias based on demographic characteristics and personal relationships. Attorneys in the survey gave the judges an average score of 1.34.

The communication measure is based on the ABA Guidelines category called "Communication Skills." The survey asks only one question here, which asks attorneys if the judge clearly explains the reasons for the decision. The average score on this measure is 1.26.

The temperament score is based on the ABA Guidelines category called "Professionalism and Temperament," but the survey asks only about judicial courtesy. The judges scored an average of 1.42.

The administrative score comes from the ABA Guidelines category called "Administrative Capacity." The survey asks attorneys about a judge's punctuality, efficiency, workload, and turnaround time. The average score on this measure was 1.38.

### *Immutable Characteristics of the Judge*

The gender and race of a judge should have no relationship to that judge's professional performance. But previous research suggests that there are significant gender and race differences on attorney survey measures of judicial performance (Burger 2007; Gill et al. 2011). This is certainly the case in the Judging the Judges survey. In this sample, 35% of the judges were women. The left side of Table 3 presents a series of difference of means tests comparing average scores for men and women. It is clear from the data that women score significantly lower than their male counterparts on each and every one of the performance categories in the Judging the Judges survey. The results show that there is a significant and sizeable difference between scores for men and women.

A similar analysis for judges of racial minority status is presented on the right side of Table 3. Unfortunately, only 5% of the judges in the sample are of a minority racial group, so it is more difficult to establish statistical significance of this difference. But the magnitude of the differences is even higher than what we saw for female judges.

### *“Objective” Measures of Judicial Performance*

Measures of judicial performance have relied very heavily on survey data like what is described above (Pelander 1998). In order to determine if the disparities outlined above are attributable to unconscious bias, it is important to consider other measures of judicial performance that may be less easily influenced by this kind of bias. There are several other ways to capture relevant performance characteristics through the use of proxy variables for legal ability, integrity, and the like. The use of these variables is critical to our ability to determine whether gender- and race-based differences in performance evaluation scores represent real differences in our sample, or alternatively if these disparities are based on systematic unconscious gender bias.

The first control variable in the model is the prestige of the judge’s legal education. This is an important cue for the respondent lawyers as to the legal ability of the judge. These lawyers will have a great deal of familiarity with the prestige of various law schools. Law school prestige is measured in “tiers” as classified by the *US News & World Reports*.<sup>3</sup> A more prestigious school is in a tier with a lower number, such that a tier one school is considered to be more prestigious than a tier two school. This means that our measure should be negatively correlated with an underlying dimension of legal ability, communication skills, and administrative capacity. Getting into a top law school requires high performance in college courses as well as strong reading, writing, and analytical skills. To make it through these higher prestige programs, students must also demonstrate organization, oral and written communication skills, and mastery of the concepts underlying the law. The average judge in the sample had a law school prestige score of 2.18, meaning that they fell somewhere in the second of the four tiers. The range of this variable, however, is from one to five. The fifth “tier” in this measure represents those judges in the sample who did not attend law school at all.

The second control variable is a measure of the judge’s reversal rate. The reversal rate is an important measure for judicial quality (Brody 2000; Posner 2000), and may capture both communication skills and legal ability. A low reversal rate means that when a judge’s decisions are reviewed, they are overturned infrequently.<sup>4</sup> Like the measure of law school prestige, a

---

<sup>3</sup> The prestige was measured as of 2010, the date of the last survey. Although this does not capture the ranking of the school at the time the judge attended, I assume this not to be problematic because there is a very high level of consistency of tier classifications over time (although rank order may vary from year to year). These scores can be found online at <http://grad-schools.usnews.rankingsandreviews.com>.

<sup>4</sup> Previous research has investigated the possibility that reversal rates are themselves a product of gender bias, but found no evidence that the decisions of female judges were appealed more often, nor that they were reversed more often (Gill et al. 2011; Walker and Barrow 1985).



lower reversal rate is associated with higher levels of judicial ability. The average value of this measure is .07, and it ranges from 0 to 43%.<sup>5</sup>

I have also included a dummy variable to indicate if the judge was appointed to the bench to fill a vacancy. Unlike most states with judicial elections, vacancies are filled by the equivalent of the merit plan. Although the evidence is mixed as to whether the merit plan selects higher quality judges than do elections, some benefits may accrue to the judge who obtains a seat on the bench without having first been elected. The most notable of these is that the judge gains incumbency status without having to stand for election (Glick 1978). Also, lawyers tend to see the merit plan as resulting in higher quality judges (O'Connor 2009), at least in part because lawyers dominate the process. In addition to this, Nevada's nonpartisan election system does not allow an opportunity for political parties to vet the judicial candidates prior to nomination, and so there may be real differences in quality between the appointed candidates and the elected ones. Previous research shows that having been first appointed tends to drive up attorney recommendations to retain judges in this institutional context (Gill et al. 2011).

A measure of judicial experience on the bench is also included in the model. We would typically expect such a measure to be positively related to judicial quality, especially as it concerns legal ability, communication skills, and administrative ability. Previous research suggests that this is the case (Epstein et al. 2003; Haire 2001). This may be in part due to the acclimation effects that show up in patterns of decision making among new judges (Brenner and Hagle 1996; Hettinger et al. 2003). The average number of years of experience in the sample is 9.05, and this ranges from 0 to 30.

The above judicial quality measures tend to capture things like legal ability, communication skills, and administrative performance. I have included a pair of measures intended to capture judicial temperament and integrity: discipline records and public scandals. The discipline variable captures the presence and outcome severity of formal disciplinary proceedings. This is an ordinal scale of increasing outcome severity.<sup>6</sup> The average is 0.56, and the range is 0 to 8. The reported scandal score is a less formal measure of judicial ethics. While it is relatively easy to file a formal complaint against a judge, many attorneys are unwilling to do so for fear of retribution (Jackson 2007). But the media generally have no such fears, and the press reports

---

<sup>5</sup> I also tried a measure of reversal rate distance from the court mean, which compares each judge's reversal rate to the mean on his or her own level of court (e.g., supreme court, district court). This variable did not perform any differently than the reversal rate measure and, as in previous research (Gill et al. 2011), this variable was insignificant in all of the models.

<sup>6</sup> This measure comes from Gill et. al (2011). The coding is as follows: 0=no complaint; 1=complaint dismissed; 2=sanctioned: required course; 3=sanctioned: required course and public apology; 4=sanctioned: public reprimand; 5=sanctioned: public reprimand and fine; 6=sanctioned: censure, required course, and fine; 7=sanctioned: removal from bench, and 8=sanctioned: removal from bench and permanently barred from holding public office in Nevada.

stories of alleged judicial impropriety with stunning frequency. Of course, the fact that such a story has been published does not mean that a judge is of lower quality, but this variable might certainly control for a judge's general reputation of integrity among local attorneys. This dummy variable was constructed to denote the presence or absence of a scandal to which the judge has been publicly connected in the main local newspaper of record, the *Las Vegas Review-Journal*. In this study, nearly a quarter of the judges have been publicly associated with some sort of personal or professional scandal.

Finally, I have included an ordinal measure of the court on which the judge is serving when the survey is conducted. This measure ranges from 1 for the municipal courts to 5 for the Nevada Supreme Court. In this sample, we have observations for a total of 39 justices from the Supreme Court, 120 district court judges, 65 family court judges, 69 judges in the justice court, and 57 in the municipal court. This measure can be thought of as a proxy measure of judicial quality because judges elected to the higher courts tend to be seen as higher in quality generally. Indeed, the judges who run for election to these higher court benches are subjected to increased public scrutiny during the campaign process.

### **Multivariate Models of Judicial Performance Evaluation Scores**

In order to determine what drives the gender- and race-based differences in judicial performance evaluation scores, a model must account for both the immutable characteristics of the judge as well as the set of objective performance measures outlined above. Because the data include repeated biennial evaluations for many of the judges in the sample, the model must account for both the non-independence of these measures and the unbalanced nature of the panel dataset. In addition, it is important to take into consideration the varying amounts of information that have gone into the creation of the judge-level aggregations. As Table 2 shows, an average of 201 attorney surveys has gone into the production of the mean scores I am using as my dependent variables. But this number ranges from 23 to 467. While it is difficult to calculate this in terms of a true response rate,<sup>7</sup> we can have more confidence in those scores that are based upon higher response rates.

For these reasons, I use a pooled weighted least squares regression model with Driscoll and Kraay (1998) standard errors. The robust standard errors are a correction for the repeated biennial measures of some of the judges (Hoechle 2007), and the model is weighted by the number of attorney responses used to create the aggregated measures. This allows us to

---

<sup>7</sup> Researchers sent out a different number of surveys each year, but the number averaged around 4,000. Fewer than 20% of attorneys responded. But attorneys were instructed to rate only those judges with which they had firsthand experience during the evaluation period, so most attorneys rated only a fraction of the judges. Because the Nevada court system does not keep records of which attorneys appear before which judges, we cannot calculate a real response rate. See Gill et. al (2011) for a discussion of this problem.

estimate our parameters relying more heavily on those observations in which we have the most confidence.<sup>8</sup>

The results of this analysis are presented in Table 4. In all, these models perform quite well. The set of independent variables explains about a third of the variance in the dependent variables. But, across the board, the coefficients on the race and gender variables are statistically significant and of high magnitude. Women score almost two tenths of a point lower than men, even after controlling for the various alternative measures of judicial performance. Minority judges fare even worse, losing more than a quarter of a point, regardless their scores on the objective measures. On a three point scale, this is of magnitude large enough to cause some alarm.

While the control variables are not the main focus of this paper, some of the interesting results cast even more doubt on the validity of the survey instrument as a measure of judicial quality. Reversal rates are not significant predictors of any of the performance measures, despite their theoretically important relationship to legal ability and communication skills. Many of the other control variables are only significant in explaining characteristics to which they are theoretically not related. For example, the law school prestige variable is not a significant predictor of legal ability, communication, or administrative skills. Having attended a prestigious law school does increase scores on integrity and judicial temperament—the two dimensions of judicial performance that seem largely unrelated to the quality of one’s legal education. Experience on the bench is negatively related to scores on all of these measures. Each additional year on the bench is associated with about a .01 point decrease in scores. Because these scores have a scale of only three points, and because these judicial careers can span decades, this can result in a high magnitude effect across a judge’s career.<sup>9</sup>

The ethics variables also present an interesting result. The presence of any news article tying the judge to a scandal works to lower the judge’s score on all of the measures by up to a point. Interestingly, the smallest magnitude effect is in the category of judicial integrity. And this isn’t due to the effects of the discipline variable. The outcome of the most severe disciplinary measures is insignificant in the model of judicial integrity scores. It does, however, serve to lower scores on administrative performance.

---

<sup>8</sup> This is a conservative approach; the models estimated without the weights showed even higher magnitude effects of gender and race on the values of the dependent variables.

<sup>9</sup> In an attempt to square this result with what the literature seem to suggest, I tried a number of alternative specifications of the relationship between experience and survey scores. I found no evidence of a nonlinear relationship, and dummy variables for “very little” or “very much” experience were also insignificant in these models.

## Discussion

Unfortunately, the results above suggest that female and minority judges have to counterbalance the unconscious bias manifested in poorly-designed surveys that acts like a thumb on the scales, systematically disadvantaging groups that have been traditionally underrepresented. It does not appear that there is a single category of questions that is susceptible to unconscious bias; the effects are significant, large, and consistent across all of the dimensions of judicial performance evaluated by the Judging the Judges survey.

Even setting the large problem of gender and race bias aside, the objective measures of judicial performance often do not have the theoretically expected effect on measures of judicial performance. The fact that reversal rates are insignificant in all of the models is more than a little worrisome. The fact that judicial discipline outcomes are unrelated to the measure of judicial integrity is also problematic. That scores go down systematically with increased experience on the bench is also troubling. Taken together, the performance of these control variables suggests that the questions in the Judging the Judges survey—and by proxy the ABA's Guidelines—do not do a good job capturing the true quality of the traits they are attempting to quantify.

But the bad news does not end there. The questions on the Judging the Judges survey are derived from the ABA Guidelines (American Bar Association 1985), and the questions in the survey are very similar to the questions used in state-sponsored JPEs around the country (Gill et al. 2011). Colorado's system of JPE is the most studied and well-funded in the nation. The head of the Institute for the Advancement of the American Legal System (IAALS) says of Colorado's survey:

All those individuals complete survey questionnaires that focus on things such as the following: was the judge prepared when he or she showed up on the bench; was the judge respectful of the people in the courtroom; did the judge move the docket along efficiently; was the judge timely in his or her rulings' and were those rulings clear and understandable? In other words, the questions focus on process, not on outcome. (Kourlis 2010, 767)

But these are *exactly* the same questions that we find on the "Judging the Judges" survey. And many of the other characteristics of the Judging the Judges survey are also common to JPEs in other states, including the primacy of the survey in the JPE process (Pelander 1998) and the low response rates (Brody 2000). As such, there is good reason to believe that the very real problems I've identified in the Judging the Judges survey are being replicated in states across the country.

None of this would be a problem, except for the fact that the results of these judicial performance evaluations actually *matter*. As Judge Kourlis explains, surveys of voters found that “if voters knew that we had a judicial performance evaluation system, they would use the information and they would trust it” (Kourlis 2010, 768). She says of these voters:

They trusted the fact that the judicial performance commissioners were looking at the right data and making good decisions; and, thus, the voters could turn to that data and those decisions to guide them. That is what an informed vote should look like. (Kourlis 2010, 768)

In short, following the ABA’s Guidelines is not enough to prevent serious bias from poisoning judicial performance evaluation scores. This is a particularly poor outcome, as it means subjecting many judges to state-sponsored evaluations that are systematically biased against women and minorities. Beyond that, the results for white male judges are similarly flawed. When formal disciplinary action is unrelated to integrity measures and law school prestige is related to temperament and integrity but not to legal knowledge and communication skills, we must consider the possibility that these surveys are not measuring what their designers had hoped they would measure. This analysis raises real questions about the overall validity of the JPE survey as a measurement of judicial performance. While Judge Kourlis laments the fact that many voters are unaware of the judicial performance evaluation data when they make their decisions (Kourlis 2010), perhaps instead we should be relieved by this fact. This does not mean that the entire enterprise of evaluating judicial performance should be abandoned; however, we must do better than this.

**Table 1: ABA Categories and Judging the Judges Questions**

<b>ABA Guidelines</b>		<b>Judging the Judges Questions</b>
<b>Legal Ability</b>		
1-1	Legal reasoning ability	The judge's rulings, whether regarding civil issues, criminal sentencing, or contempt are appropriate.
1-2	Knowledge of substantive law	The judge demonstrates familiarity with the case record and documents, and fairly weighs all evidence and arguments before rendering a decision
1-3	Knowledge of rules of procedure and evidence	The judge properly applies the law, rules of procedure, and rules of evidence.
<b>Integrity and Impartiality</b>		
2-1	Avoidance of impropriety and the appearance of impropriety.	The judge's professional conduct is free from impropriety and the appearance of impropriety.
2-3	Absence of favor or disfavor toward anyone, including but not limited to favor or disfavor based upon race, sex, religion, national origin, disability, age, sexual orientation, or socioeconomic status.	The judge's conduct is free from bias on the basis of race or ethnic origin. The judge's conduct is free from bias on the basis of gender. The judge's conduct is free from bias on the basis of religion.
2-6	Basing decisions on the law and the facts without regard to the identity of the parties or counsel, and with an open mind in considering all issues.	The judge's conduct is free from bias on the basis of parties or attorneys involved in the action.
<b>Communication Skills</b>		
3-1	Clear and logical oral communication while in court.	The judge clearly explains the basis for his or her decisions.
3-2	Clear and logical written decisions.	
<b>Professionalism and Temperament</b>		
4-2	Treating people with courtesy	The judge is courteous.
<b>Administrative Capacity</b>		
5-1	Punctuality and preparation for court.	The judge is punctual in convening court keeps business moving, and does an amount of work fair to taxpayers and other judges.
5-4	Making decisions and rulings in a prompt, timely manner.	The judge issues orders, judgments, decrees, or opinions without unnecessary delay.

**Table 2: Descriptive Statistics**

Variable	Mean	Std. Dev.	Min	Max
<i>Dependent Variables</i>				
Legal Ability	1.24	0.27	0.23	1.75
Integrity	1.34	0.27	0.35	1.78
Communication	1.26	0.28	0.27	1.77
Temperament	1.42	0.32	0.18	1.92
Administrative	1.38	0.24	0.26	1.81
<i>Independent Variables</i>				
Minority Status	0.05	0.23	0.00	1.00
Female	0.35	0.48	0.00	1.00
Law School Prestige	2.18	1.24	1.00	5.00
Reversal Rate	0.07	0.08	1.00	0.43
Appointed	0.40	0.49	0.00	1.00
Years Experience	9.05	6.84	0.00	30.00
Discipline	0.56	1.51	0.00	8.00
Reported Scandal	0.23	0.42	0.00	1.00
Level of Court	2.91	1.24	1.00	5.00
<i>Weight Variable</i>				
Responses	200.90	104.21	23.00	467.00

**Table 3: Mean Difference by Sex & Minority Racial Status of Judge**

		Mean	T	p		Mean	T	p
Legal Ability								
	Male	1.28			White	1.23		
	Female	1.09			Non-White	1.02		
	Difference	0.19	3.36***	.001	Difference	0.21	1.73*	.044
Integrity								
	Male	1.36			White	1.33		
	Female	1.26			Non-White	1.15		
	Difference	0.10	1.89*	.031	Difference	0.19	1.67*	.048
Communication								
	Male	1.29			White	1.26		
	Female	1.14			Non-White	1.01		
	Difference	0.15	2.85**	.003	Difference	0.24	2.13*	.018
Temperament								
	Male	1.46			White	1.42		
	Female	1.31			Non-White	1.23		
	Difference	0.16	2.32**	.011	Difference	0.19	1.35	.090
Administrative								
	Male	1.42			White	1.37		
	Female	1.25			Non-White	1.18		
	Difference	0.17	3.61***	.000	Difference	0.20	1.85*	.034



**Table 4: Weighted Pooled OLS Models of Evaluation Scores**

	Legal Ability	Integrity	Communication	Temperament	Administrative
Minority Status	-0.269*** (.023)	-0.252*** (.052)	-0.353*** (.033)	-0.410*** (.071)	-0.199*** (.010)
Female Judge	-0.193*** (.049)	-0.109** (.035)	-0.142** (.043)	-0.171*** (.051)	-0.205*** (.054)
Law School Prestige	-0.006 (.006)	-0.025*** (.006)	-0.006 (.011)	-0.034*** (.004)	.008 (.006)
Reversal Rate	0.083 (.097)	0.017 (.195)	0.181 (.182)	0.007 (.202)	-0.078 (.101)
First Appointed	0.042* (.019)	0.080** (.029)	0.052 (.032)	0.068** (.026)	0.004 (.012)
Years on Bench	-.009** (.003)	-0.012*** (.002)	-.009*** (.002)	-0.010** (.003)	-0.005*** (.001)
Discipline	-0.020 (.015)	-0.016 (.013)	-0.014 (.011)	-0.015 (.015)	-0.045*** (.012)
Reported Scandal	-0.113*** (.013)	-0.068** (.023)	-0.133*** (.017)	-0.160*** (.023)	-0.092*** (.010)
Court Level	0.037** (.014)	0.069* (.029)	0.053* (.026)	0.011 (.008)	-0.029** (.010)
Constant	1.335*** (.055)	1.370*** (.131)	1.283*** (.131)	1.653*** (.064)	1.636*** (.049)
Obs.	344	344	344	304	332
Judges	93	93	93	86	92
r <sup>2</sup>	.30	.33	.29	.35	.37
F	382***	691 ***	786***	79***	332***
Root MSE	4.8367	5.2733	4.8976	5.6133	5.0484

These are Weighted Pooled OLS models with Driscoll-Kraay standard errors. The weight in this model is the number of attorney responses for each evaluation. n=343 observations across 93 judges with two exceptions: The Judging the Judges survey did not collect temperament scores for judges sitting on the Nevada Supreme Court, and they did not collect administrative performance scores for judges sitting on the Nevada Supreme Court in 1998.

## Works Cited

- American Bar Association. 1985. "Guidelines for the Evaluation of Judicial Performance." Washington, D.C.
- . 2005. "Guidelines for the Evaluation of Judicial Performance with Commentary." Chicago, IL.
- Bartlett, F. C. 1932. *Remembering*. New York: MacMillan.
- Bowman, Cynthia Grant. 1998. "Bibliographical Essay: Women and the Legal Profession." *American University Journal of Gender, Social Policy and Law* 7:149.
- Brenner, Saul, and Timothy M. Hagle. 1996. "Opinion Writing and Acclimation Effects." *Political Behavior* 18 (3):235-61.
- Brody, David C. 2000. "Judicial Performance Evaluations by State Governments: Informing the Public While Avoiding the Pitfalls." *Justice System Journal* 21:333-56.
- Burger, Gary K. 2007. "Attorney's Ratings of Judges: 1998-2006." Mound City, MO: Report to the Mound City Bar.
- Cann, Damon. 2006. "Beyond Accountability and Independence: Judicial Selection and State Court Performance." *Judicature* 90:226.
- Carnes, Molly, Stacie Geller, Eve Fine, Jennifer Sheridan, and Jo Handelsman. 2005. "NIH Director's Pioneer Awards: Could the Selection Process Be Biased against Women?" *Journal of Women's Health* 14 (8):684-92.
- Choi, Stephen J., Mitu Gulati, and Eric A. Posner. 2009. "Judicial Evaluations and Information Forcing: Ranking State High Courts and their Judges." *Duke Law Journal* 58:1313.
- Cleeremans, A. 2003. "Implicit Learning Models." In *Encyclopedia of Cognitive Science*, ed. L. Nadel. New York: Nature Publishing Group.
- Clydesdale, Timothy T. 2004. "A Forked River Runs Through Law School: Toward Understanding Race, Gender, Age, and Related Gaps in Law School Performance and Bar Passage." *Law & Social Inquiry* 29 (4):711-69.
- Coontz, Phyllis D. 1995. "Gender Bias in the Legal Profession: Women 'See' It, Men Don't." *Women & Politics* 15 (2):1-22.
- Driscoll, John C., and Aart C. Kraay. 1998. "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data." *The Review of Economics and Statistics* 80 (4):549-60.
- Durham, Christine M. 2000. "Gender and Professional Identity: Unexplored Issues in Judicial Performance Evaluation." *Judges' Journal* 39 (2):13-6.
- Epstein, Lee, Jack Knight, and Andrew D. Martin. 2003. "The Norm of Prior Judicial Experience and Its Consequences for Career Diversity on the U.S. Supreme Court." *California Law Review* 91 (4):903-65.
- Esterling, Kevin M. 1998. "Judicial Accountability the Right Way." *Judicature* 82:206.
- Fiske, Susan T., Donald N. Bersoff, Eygene Borgida, Kay Deaux, and Madeline E. Heilman. 1991. "Social Science Research on Trial: Use of Sex Stereotyping Research in *Price Waterhouse v. Hopkins*." *American Psychologist* 46 (10):1049-160.
- Geary, Frank. 2008. "Lawyers Rate Female Jurists as Less Courteous than Men." *Las Vegas Review-Journal*, May 21.
- Gill, Rebecca D., Sylvia R. Lazos, and Mallory M. Waters. 2011. "Are Judicial Performance Evaluations Fair to Women and Minorities? A Cautionary Tale from Clark County, Nevada." *Law & Society Review* 45 (3):731-59.
- Glick, Henry R. 1978. "The Promise and the Performance of the Missouri Plan: Judicial Selection in the Fifty States." *University of Miami Law Review* 32 (3):509-43.

- Gordon, J. Cunyon. 2003. "Painting by Numbers: "And, Um, Let's Have a Blac Lawyer Sit at Our Table." *Fordham Law Review* 71 (4):1257.
- Haire, Susan B. 2001. "Rating the Ratings of the American Bar Association Standing Committee on Federal Judiciary." *Justice System Journal* 22 (1):1-18.
- Hall, William K. 1985. "Judicial Retention Elections: Do Bar Association Polls Increase Voter Awareness?." Urbana, IL: Institute of Government and Public Affairs, University of Illinois.
- Heilman, Madeline E. 1983. "Sex Bias in Work Settings: The Lack of Fit Model." *Research in Organizational Behavior* 5:269-98.
- Hekman, David R., Karl Aquino, Bradley P. Owens, Terence R. Mitchell, Pauline Schilpzand, and Keith Leavitt. 2010. "An Examination of Whether and How Racial and Gender Biases Influence Customer Satisfaction." *Academy of Management Review* 35 (2):238-64.
- Hettinger, Virginia A., Stefanie A. Lindquist, and Wendy L. Martinek. 2003. "Acclimation Effects and Separate Opinion Writing on the United States Courts of Appeals." *Social Science Quarterly* 84 (4):792.
- Hoechle, Daniel. 2007. "Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence." *Stata Journal* 7 (3):281-312.
- Jackson, Jeffrey. 2007. "Beyond Quality: First Principles in Judicial Selection and their Application to a Commission-Based Selection System." *Fordham Urban Law Journal* 34:125.
- Kang, Jerry, and Mahzarin Banaji. 2006. "Fair Measures: A Behavioral Realist Revision of "Affirmative Action"." *California Law Review* 94:1063.
- Kourlis, Rebecca. 2010. "Judicial Performance Evaluation." *Wayne Law Review* 56:765.
- Krieger, Linda Hamilton. 1995. "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity." *Stanford Law Review* 47 (6):1161-248.
- Lee, Audrey J. 2005. "Unconscious Bias Theory in Employment Discrimination Litigation." *Harvard Civil Rights-Civil Liberties Law Review* 40 (481-503):481.
- O'Connor, Sandra Day. 2009. "Judicial Independence and Civic Education." *Utah Bar Journal* 22 (5):10-9.
- Pelander, A. John. 1998. "Judicial Performance Review in Arizona: Goals, Practical Effects and Concerns." *Arizona State Law Journal* 30:643.
- Posner, Richard A. 2000. "Is the Ninth Circuit Too Large? A Statistical Study of Judicial Quality." *Journal of Legal Studies* 29 (2):711-9.
- Rhode, Deborah L. 2001. "The Unfinished Agenda: Women and the Legal Profession." Chicago, IL: ABA Commission on Women in the Profession.
- Sterling, Joyce S. 1993. "The Impact of Gender Bias on Judging: Survey of Attitudes Toward Women Judges." *Colorado Law Review* 22:257.
- Walker, Thomas G., and Deborah J. Barrow. 1985. "The Diversification of the Federal Bench: Policy and Process Ramifications." *Journal of Politics* 47:596-617.
- White, Penny J. 2009. "Retention Elections in a Merit-Selection System: Balancing the Will of the Public with the Need for Judicial Independence and Accountability: Using Judicial Performance Evaluations to Supplement Inappropriate Voter Cues and Enhance Judicial Legitimacy." *Missouri Law Review* 74:635.
- Wood Gill, Rebecca, and Sylvia R. Lazos. 2009. "Reflections in Response to the Nevada Judicial Education Pilot Project." In *William S. Boyd School of Law Research Papers*. Las Vegas, NV.